Enhanced EHACBLalign - GCN Method for Protein Remote Homology Detection and Fold Recognition

Gopinath Krishnaraj[1, 2], Rajendran Gurusamy[3]

[1]Research Scholar, [2]Assistant Professor, [3]Associate Professor and Head
[1]Department of Computer Science, Periyar University, Salem– 636 011, Tamil Nadu, India
[2]Department of Computer Applications, Sona College of Arts and Science, Salem – 636 005, Tamil Nadu, India
[3]Department of Computer Science, Govt. Arts and Science College, Modakkurichi, Erode – 638104, Tamil Nadu, India

Corresponding author: Gopinath Krishnaraj

Abstract

Protein Remote Homology Detection and Fold Recognition is a fundamental task in bioinformatics, essential for understanding protein functions, facilitating drug discovery, and annotating genes. Traditional approaches, such as Convolutional Neural Networks (CNNs), often face challenges in processing the vast and complex data associated with protein sequences, leading to difficulties in accurately recognizing protein homologies. This paper introduces a novel approach leveraging Graph Convolutional Networks (GCNs) to address this challenge and to prune uninformative edges within the graph, effectively reducing noise and enhancing the accuracy and efficiency of homology detection. By embedding protein sequences into a vector space and using a Softmax classifier for final classification, the GCN method captures intricate relationships among proteins, resulting in superior performance compared to existing methods.

The proposed method's effectiveness is validated through extensive experiments on benchmark datasets, including SCOP 1.53, SCOP 1.67, and superfamily datasets, demonstrating significant improvements in prediction in terms of accuracy, precision, recall and F-measure. The findings expose the potential of GCNs in Remote Homology Detection and Fold Recognition.

Keywords: **Remote Homology Detection and Fold Recognition, Graph Convolutional Networks (GCN), Softmax Classifier**

1. Introduction

Remote Homology Detection and Fold Recognition is a cornerstone of bioinformatics and computational biology, playing a critical role in deciphering the evolutionary relationships among proteins. These relationships are vital for numerous applications; including protein function prediction, drug discovery, and gene annotation, all of which have profound implications in both scientific research and the pharmaceutical industry. As biological data continues to expand exponentially, the challenge of accurately recognizing protein homologies from vast and complex protein sequences has become increasingly pronounced [1].

Traditional methods for Remote Homology Detection and Fold Recognition, particularly those utilizing Convolutional Neural Networks (CNNs), have shown significant promise due to their ability to automatically learn features from raw data, a key advantage over earlier hand-crafted feature approaches [2]. However, despite their success in various domains, CNNs are not ideally suited for handling the non-Euclidean nature of biological data, such as protein sequences, which are better represented as graphs. The fixed-size grid structure inherent in CNNs limits their capability to capture the intricate, non-linear relationships that characterize biological data, leading to inefficiencies in learning and difficulties in accurately recognizing remote protein homologies [3].

The complexity of Remote Homology Detection and Fold Recognition is further compounded by the fact that protein sequences often contain a vast amount of uninformative or redundant data. CNNs, when applied to such large-scale sequences, may struggle to effectively filter out this noise, which can result in reduced prediction accuracy and increased computational costs [4]. This limitation is particularly evident in the context of large-scale datasets, where the processing of redundant features can lead to significant inefficiencies in the learning process [5].

To address these challenges, this research introduces a novel approach based on Graph Convolutional Networks (GCNs). GCNs are a type of neural network specifically designed to operate on graph-structured data, making them exceptionally well-suited for tasks involving the analysis of complex relationships within biological networks [6]. Unlike CNNs, GCNs are inherently capable of handling the non-Euclidean nature of biological data, which allows them to capture more complex and meaningful relationships among protein sequences. This capability is particularly valuable in the context of Remote Homology Detection and Fold Recognition, where understanding the intricate connections between proteins is crucial for accurate homology detection [7].

In this research, the GCN method is further enhanced by integrating a Softmax classifier, which is used to classify proteins based on their embedding in a vector space. In this vector space, proteins that share similar characteristics are positioned closer together, facilitating more accurate and efficient homology detection [8]. This approach not only improves prediction accuracy but also accelerates the convergence of the model and reduces computational overhead, making it more suitable for large-scale bioinformatics applications [9].

The proposed GCN-based method is validated through extensive experiments on well-established benchmark datasets, including SCOP 1.53, SCOP 1.67, and superfamily datasets. These experiments demonstrate the superiority of the GCN model over traditional CNN-based approaches, particularly in terms of evaluation metrics of computational efficiency [10]. The findings suggest that GCNs, with their ability to control the significant advancement in the field of Remote Homology Detection and Fold Recognition [11].

2. Literature Review

The recognition of remote protein homologies is a basic challenge in bioinformatics, crucial for understanding evolutionary relationships, predicting protein functions, and advancing drug discovery efforts. Over the years, various computational methods have been developed to tackle this problem, each with its strengths and limitations.

The Basic Local Alignment Search Tool (BLAST) marked a significant milestone in sequence analysis. BLAST's ability to quickly identify regions of similarity between sequences has made it a widely used tool for detecting homologous proteins. Its underlying algorithm relies on pairwise sequence

alignment, where sequences are compared directly to identify similar regions that may indicate evolutionary or functional relationships [12]. However, BLAST's approach is less effective for identifying remote homologs, especially when sequence similarity is low due to extensive evolutionary divergence.

To address the limitations of direct sequence alignment methods like BLAST and Hidden Markov Models (HMMs), a more advanced approach was developed by modeling the conserved regions of protein families probabilistically. This allows the detection of remote homologs even when sequence similarities are not immediately apparent through direct comparison [13]. HMMs have become a critical tool in bioinformatics, particularly for tasks that involve detecting protein families and identifying functional motifs within sequences.

Despite the advancements offered by HMMs, the growing complexity and scale of biological data necessitated the exploration of more powerful computational techniques. Machine learning methods, particularly Support Vector Machines (SVMs), were introduced to enhance protein classification tasks and incorporating pairwise sequence similarity scores into an SVM framework, the detection of remote protein homologies could be significantly improved [14]. SVMs provided a flexible and robust way to classify proteins, leveraging the ability to learn from complex and high-dimensional data. However, traditional machine learning approaches like SVMs often struggle to fully capture the intricate relationships within large-scale biological datasets, particularly when the data is represented in non-Euclidean spaces such as graphs.

The advent of deep learning further revolutionized the field, with Convolutional Neural Networks (CNNs) being particularly effective for tasks involving the extraction of features from biological sequences. CNNs have been applied to Remote Homology Detection and Fold Recognition with considerable success. It was proved that CNNs could automatically learn hierarchical features from protein sequences, leading to improved recognition of remote homologs [15]. However, CNNs are inherently limited by their grid-based architecture, which is not well-suited for the non-linear and irregular structure of biological data. This limitation is particularly problematic in the context of protein sequences, where the relationships between amino acids are better represented as graphs rather than as linear or grid-like structures.

To overcome the constraints of CNNs, researchers have increasingly turned to Graph Convolutional Networks (GCNs). GCNs are designed to operate directly on graph-structured data, making them ideal for tasks where the data's inherent structure is non-Euclidean [16]. GCNs have shown significant promise in a variety of applications, including Remote Homology Detection and Fold Recognition, where they can model the complex relationships between proteins more effectively than traditional deep learning methods. By leveraging the graph structure, GCNs can capture both local and global patterns within the data, leading to more accurate predictions of remote homologies.

The development of graph embedding techniques has further enhanced the utility of GCNs in bioinformatics. Graph embedding methods transform graph-structured data into continuous vector representations, which can then be used for various downstream tasks such as classification and clustering. Node2Vec algorithm is a prime example of this approach, offering a scalable method for learning feature representations from graph data by optimizing a trade-off between breadth-first and depth-first search strategies [17]. In the context of Remote Homology Detection and Fold Recognition, graph embedding allow for the effective representation of protein sequences in a way that captures their structural and functional relationships, facilitating more accurate homology predictions.

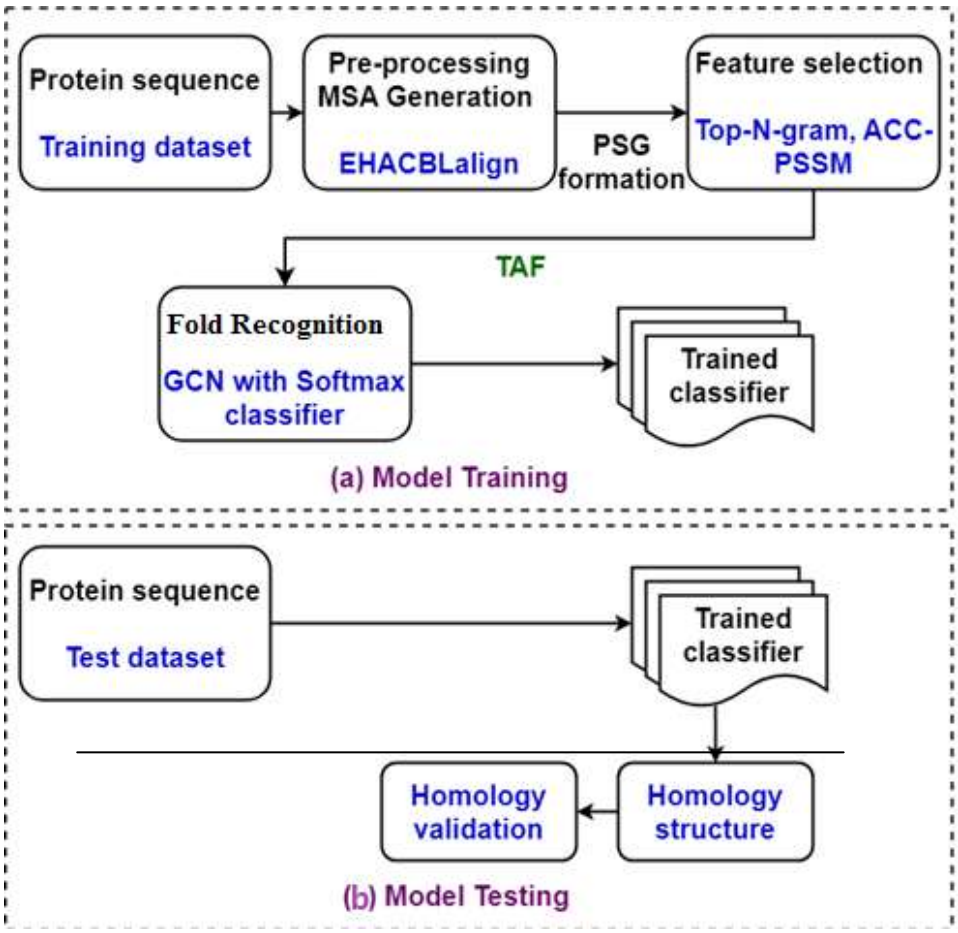
The integration of graph attention mechanisms with GCN represents another significant advancement in the field. Graph Attention Networks (GAT), which employ attention mechanisms to differentially weight the contributions of neighboring nodes during the aggregation process [18]. This selective focus on the most informative connections enhances the ability of GCNs to capture complex relationships in biological data, further improving the accuracy of Remote Homology Detection and Fold Recognition.

The integration of GCN with other deep learning models, such as auto encoders and variation auto encoders, has opened new avenues for Remote Homology Detection and Fold Recognition research [19]. These hybrid models combine the strengths of GCN in handling graph-structured data with the generative capabilities of auto encoders, leading to more accurate and robust homology predictions [20]. On graph-based methods for protein structure alignment exemplifies this approach, highlighting the potential of hybrid models to address the challenges of Remote Homology Detection and Fold Recognition [21].

3. Methodology of Proposed work

In this section, the EHACBLalign-GCN method for Remote Homology Detection and Fold Recognition is explained briefly. A block diagram of the EHACBLalign-GCN method is presented in Figure 1.

Figure 1. Block Diagram of EHACBLalign-GCN method



After generating MSAs using EHACBLalign and extracting TAF (Top-N-gram, ACC-PSSM, Features), the GCN followed by a softmax classifier is used for Remote Homology Detection and Fold Recognition.

3.1. Design of EHACBLalign - GCN Model

The Remote Homology Detection and Fold Recognition using MSAs followed by graph learning model comprise two stages. In stage 1, a Protein Similarity Graph (PSG) is created, whereas in Phase 2, the protein remote homologs on the graph generated are recognized.

Stage 1: Generate a graph depending on predefined thresholds on alignment scores: It involves constructing a network where nodes represent protein sequences, and edges indicate significant alignment scores between them. The alignment scores, which measure the degree of similarity between protein sequences, are first computed using a sequence alignment tool or algorithm. These scores are then compared against a predefined threshold to determine whether an edge should be created between two nodes. (i) If the alignment score between two sequences exceeds the threshold, an edge is drawn, signifying a meaningful similarity or potential homology. This threshold-based graph generation allows the construction of a protein similarity network that highlights only the most relevant connections, thereby reducing noise and enhancing the clarity of relationships within large biological datasets. The resulting graph is a sparse representation of the protein sequences, emphasizing only those relationships that meet the criteria for significant similarity, which is crucial for downstream tasks such as clustering, classification, or further analysis in bioinformatics. (ii) Prune the graph to preserve informative edges: Pruning helps to enhance the recognition accuracy. In this stage, the following different methods are applied to induce sparsity in the Protein Similarity Graph by pruning uninformative edges.

Stage 2: Categorize nodes on the sparse graph: After obtaining the sparse PSG, the embedding of the nodes is obtained by the unsupervised graph learning model for Remote Homology Detection and Fold Recognition. The processes in this model are unsupervised graph clustering, unsupervised node embedding, and semi-supervised GCN. In a graph clustering technique, the PSG is grouped into many clusters based on the Markov clustering scheme with every cluster defining a protein remote homologs. Because this clustering does not utilize ground truth classes (protein homologs), it is called an unsupervised scheme. In the unsupervised node embedding technique, nodes are embedded to a d-dimensional vector space using the DeepWalk method such that connected nodes on the graph remain close in the embedding space. After embedding of nodes is acquired, a softmax classifier is trained to recognize the protein remote homologs according to node embedding. At last, the semi-supervised GCN method requires ground truth homologs for a subset of proteins. According to the ground truth labels for a subset of vertices, the GCN model is trained and utilized to recognize protein remote homologs for other nodes in the test data.

4. Experimental Results

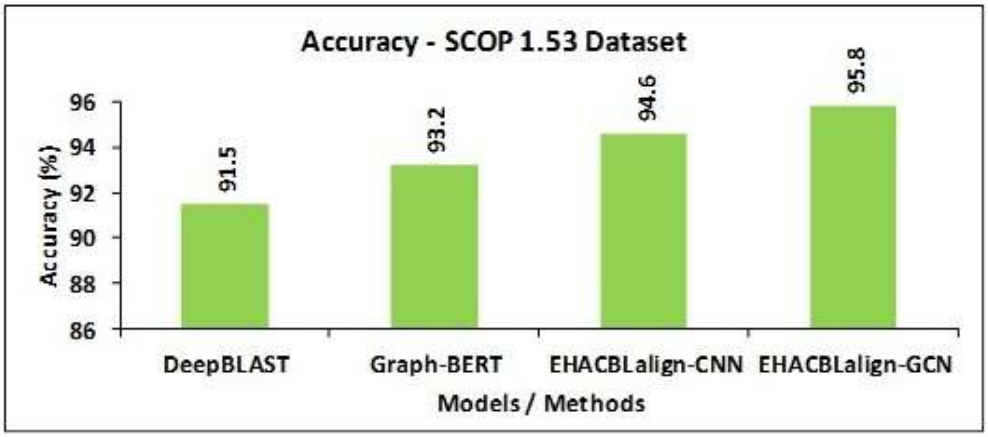
The performance of the proposed EHACBLalign-GCN method is evaluated with the existing models such as the DeepBLAST, Graph-BERT and EHACBLalign-CNN in the MATLAB 2019b using three benchmark datasets, namely the SCOP 1.53, SCOP 1.67, and the superfamily corpus are acquired to assess the effectiveness of selective MSA algorithms. The SCOP 1.53 dataset possesses 4532 PSs from 54 groups, whilst the SCOP 1.67 dataset possesses 11037 PSs from 102 groups. The superfamily dataset possesses 1195 folds of 1962 superfamilies. A superfamily is a corpus that contains labels for each PS's morphological properties. Depending on a collection of HMMs that represent structural protein motifs at the tier of the SCOP superfamily, it was built. The labels are produced by matching PSs from approximately 2478 fully sequenced genomes to HMMs.

4.1 Accuracy

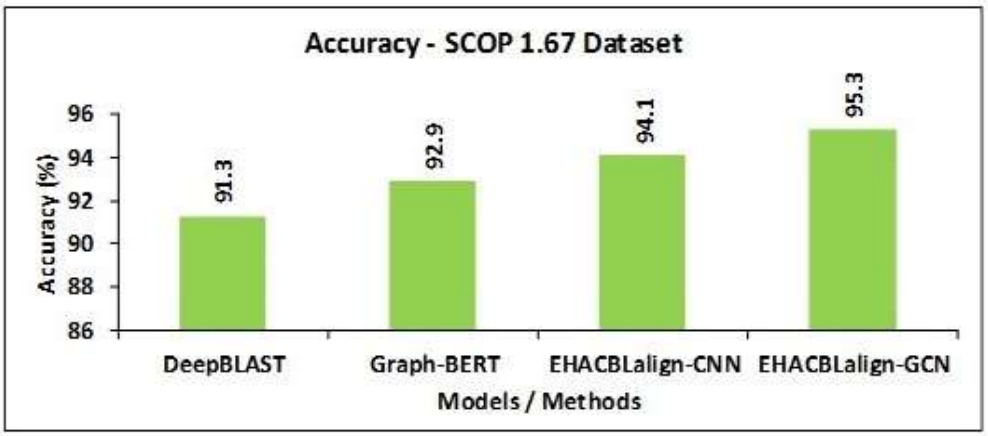
The accuracy is calculated using

$$Acc = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{TP + TN + FP + False\ Negative\ (FN)} \quad (1)$$

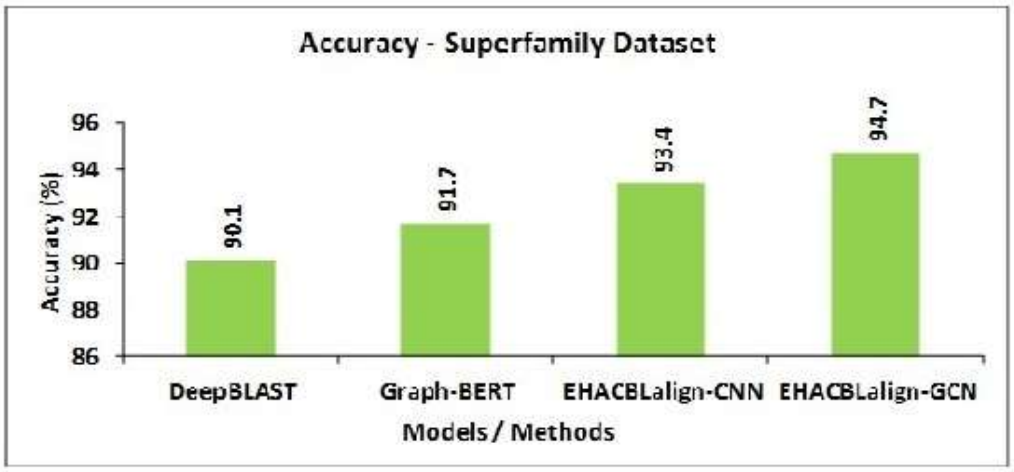
Figure 2. Comparison of Accuracy for EHACBLalign – GCN Method against Existing Models / Methods using SCOP 1.53, SCOP 1.67 and Superfamily Datasets



(a)



(b)



(c)

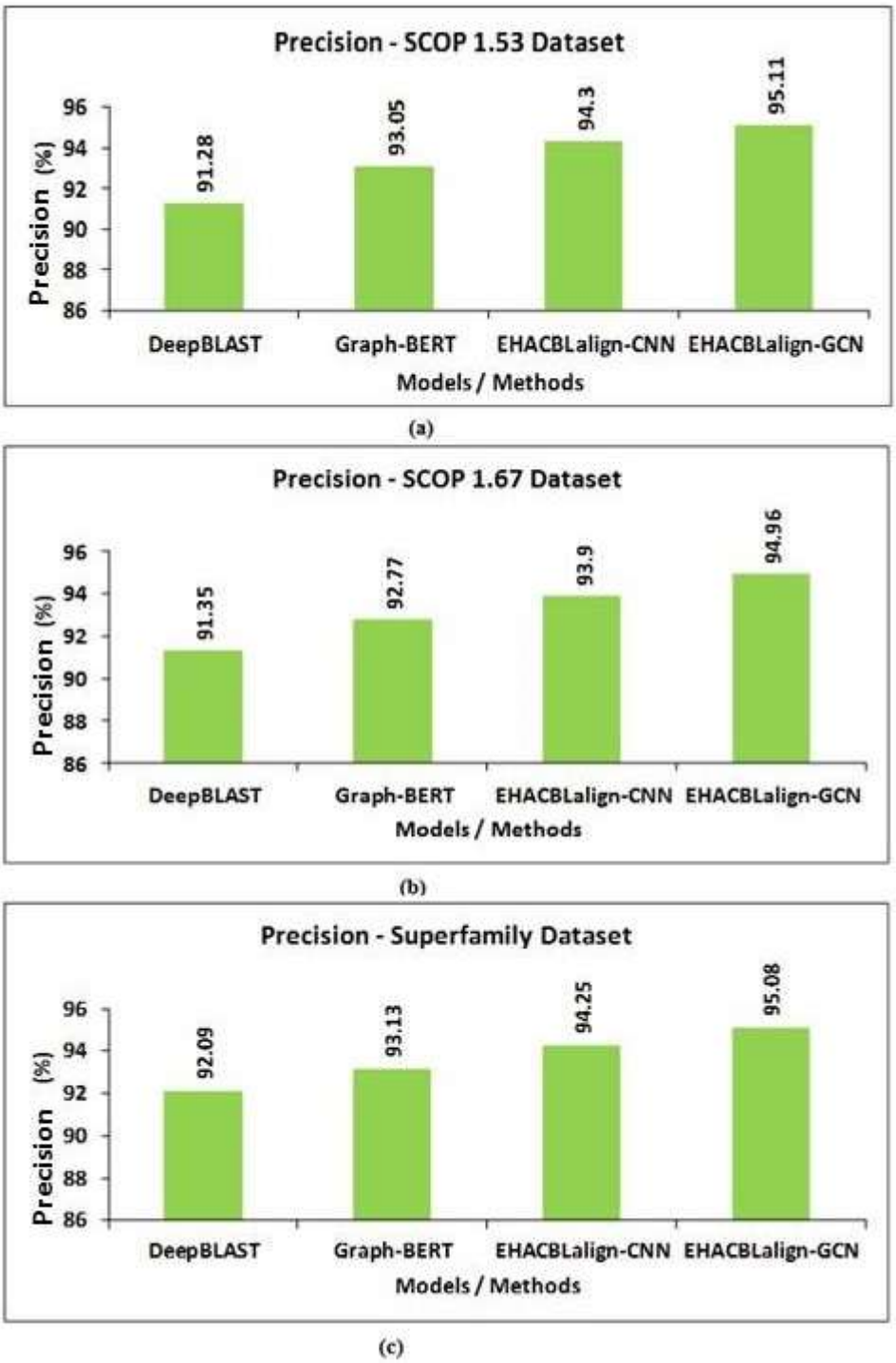
The accuracy of EHACBLalign – GCN is 95.8%, 95.3% and 94.7% respectively for SCOP 1.53, SCOP 1.67 and Superfamily datasets which are higher when compared with well-known models / methods as shown in Figure 2.

4.2 Precision

It specifies the percentage of perfectly aligned locations.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Figure 3. Comparison of Precision for EHACBLalign-GCN Method against Existing Models / Methods using SCOP 1.53, SCOP 1.67 and Superfamily Datasets



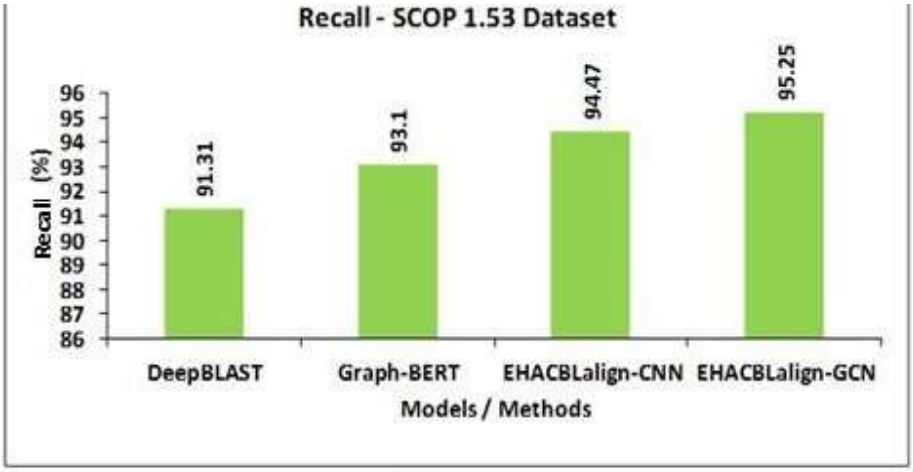
The precision of EHACBLalign – GCN is 95.11%, 94.96% and 95.08% respectively for SCOP 1.53, SCOP 1.67 and Superfamily datasets which are higher when compared with well-known models / methods as shown in Figure 3.

4.3 Recall

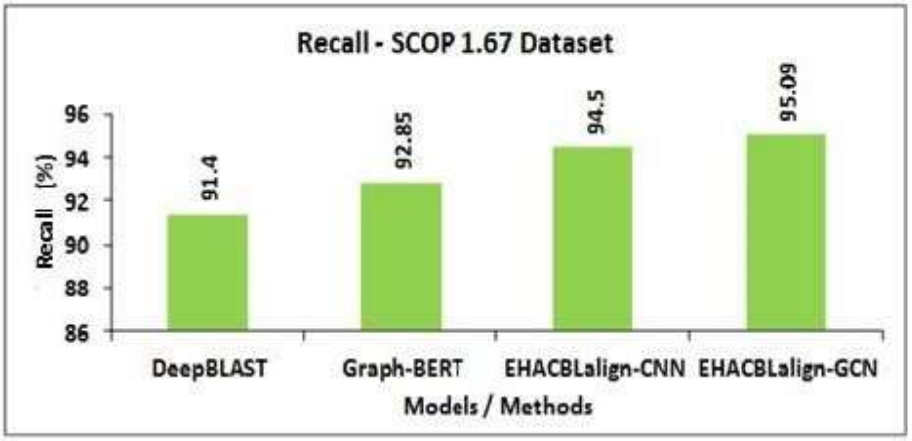
It specifies the proportion of precisely aligned residues among those that are aligned.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

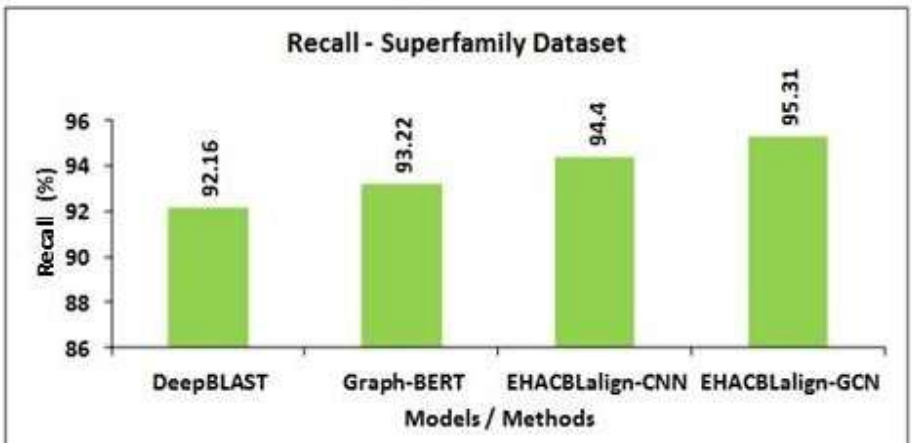
Figure 4. Comparison of Recall for EHACBLalign-GCN Method against Existing Models / Methods using SCOP 1.53, SCOP 1.67 and Superfamily Datasets



(a)



(b)



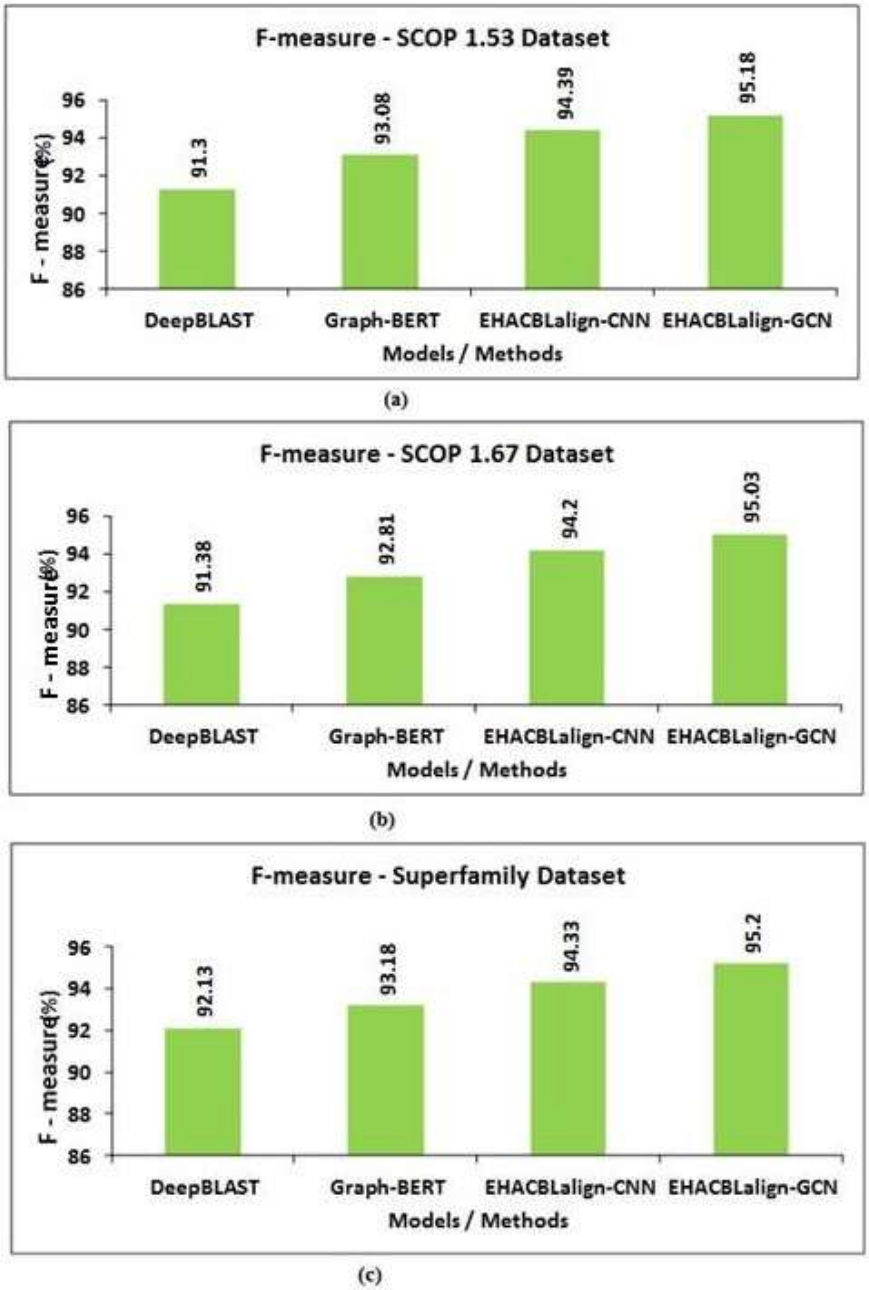
The recall of EHACBLalign – GCN is 95.25%, 95.09% and 95.31% respectively for SCOP 1.53, SCOP 1.67 and Superfamily datasets which are higher when compared with well-known models / methods as shown in Figure 4.

4.4 F-measure

It defines the F-measure of proposed and existing Protein Remote Homology Detection and Fold Recognition techniques

$$\text{F measure} = \frac{2 \text{ x Precision x Recall}}{\text{precision} + \text{recall}} \quad (4)$$

Figure 5. Comparison of f-measure for EHACBLalign-GCN Method against Existing Models / Methods using SCOP 1.53, SCOP 1.67 and Superfamily Datasets



The F-measure of EHACBLalign – GCN is 95.18%, 95.03% and 95.2% respectively for SCOP 1.53, SCOP 1.67 and Superfamily datasets which are higher when compared with well-known models / methods as shown in Figure 5.

4.5 ROC and ROC50

A comparison of ROC and ROC50 values for the EHACBLalign-GCN method against DeepBLAST, Graph-BERT, and EHACBLalign-CNN models on the SCOP 1.53 dataset is demonstrated in Figure 6(a) and 6(b).

Figure 6 (a) Comparison of ROC for EHACBLalign-GCN Method against Existing Models / Methods using SCOP 1.53 Dataset

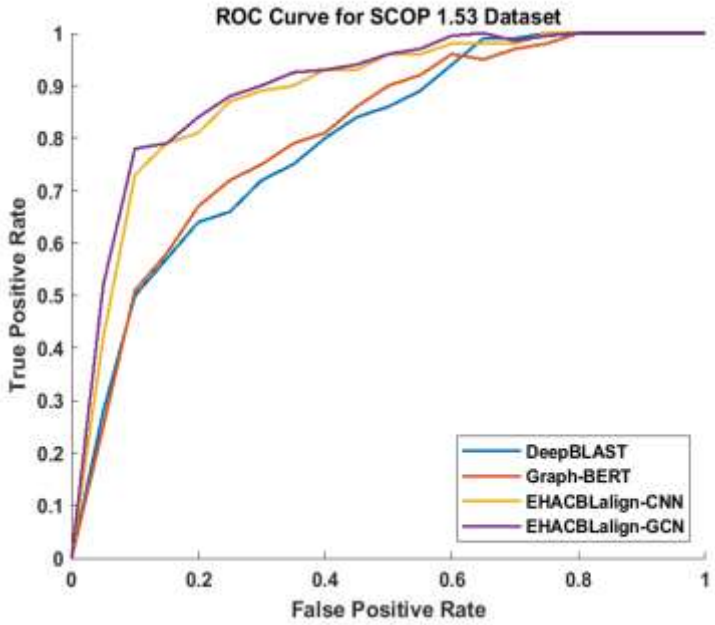
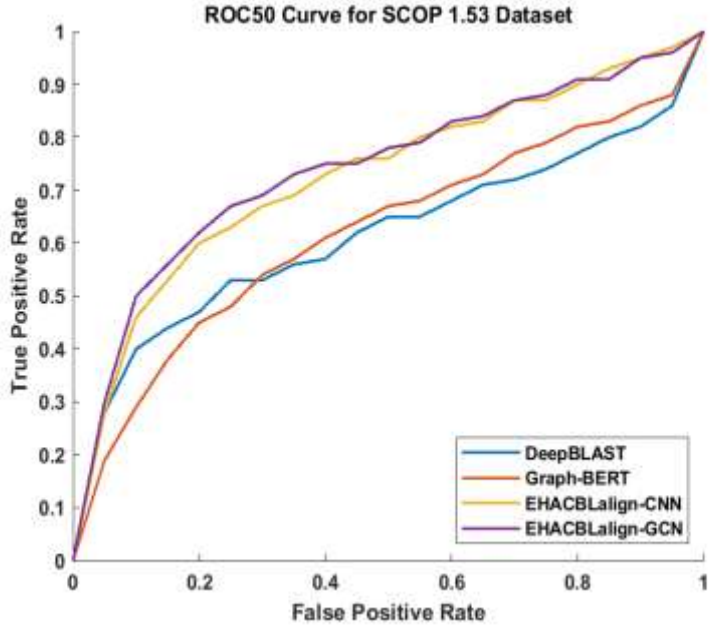


Figure 6 (b) Comparison of ROC50 for EHACBLalign-GCN Method against Existing Models / Methods using SCOP 1.53 Dataset



It can be addressed that the ROC of the EHACBLalign-GCN method is increased by 4.91%, 2.33%, and 0.5% compared to the DeepBLAST, Graph-BERT, and EHACBLalign-CNN models, respectively. Also, the ROC50 of the EHACBLalign-GCN method is 6.49%, 2.66%, and 1.1% greater than the DeepBLAST, Graph-BERT, and EHACBLalign-CNN models, respectively. This realizes that the EHACBLalign-GCN method can enhance the Remote Homology Detection and Fold Recognition performance in contrast with the other models on the SCOP 1.53 dataset.

Figure 7 (a). Comparison of ROC for EHACBLalign-GCN Method against Existing Models/ Methods using SCOP 1.67 Dataset

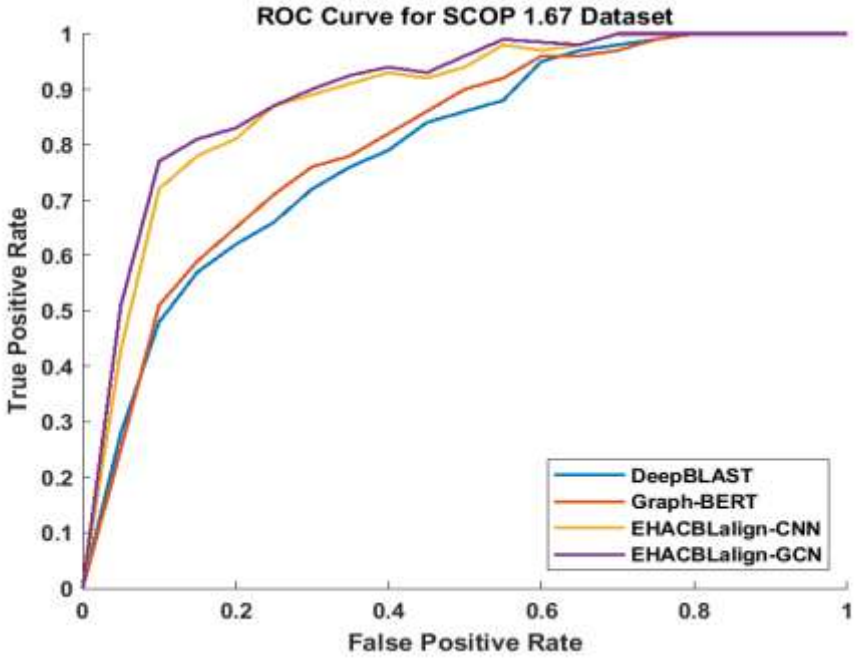
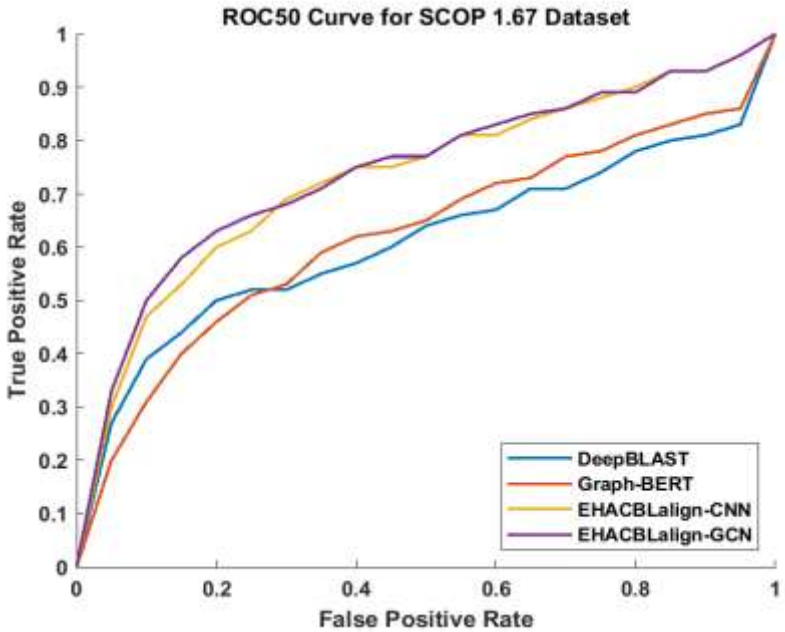


Figure 7(b). Comparison of ROC50 for EHACBLalign-GCN Method against Existing Models/ Methods using SCOP 1.67 Dataset



A comparison of ROC and ROC50 values for the EHACBLalign-GCN method against DeepBLAST, Graph-BERT, and EHACBLalign-CNN models on the SCOP 1.67 dataset is shown in Figure 7(a) and 7(b). It can be noted that the ROC of the EHACBLalign-GCN method is increased by 3.33%, 2.03%, and 1.01% compared to the DeepBLAST, Graph-BERT, and EHACBLalign-CNN models, respectively. Also, the ROC50 of the EHACBLalign-GCN model is 5.63%, 3.4%, and 1.36% greater than the DeepBLAST, Graph-BERT, and EHACBLalign-CNN models, respectively. This indicates that the EHACBLalign-GCN method can improve the Remote Homology Detection and Fold Recognition performance compared to the other models on the SCOP 1.67 dataset.

Figure 8(a) and 8(b) plots the ROC and ROC50 values of the EHACBLalign-GCN model against DeepBLAST, Graph-BERT, and EHACBLalign-CNN models on the superfamily dataset. It can be seen that the ROC of the EHACBLalign-GCN method is 4.31%, 2.38%, and 1.01% higher than the DeepBLAST, Graph-BERT, and EHACBLalign-CNN models, respectively. Also, the ROC50 of the EHACBLalign-GCN model is 4.05%, 2.5%, and 0.78% superior to the DeepBLAST, Graph-BERT, and EHACBLalign-CNN models, respectively. This shows that the EHACBLalign-GCN method boosts the efficiency of recognizing the protein remote homologs compared to the other models on the superfamily dataset.

Figure 8(a). Comparison of ROC for EHACBLalign-GCN Method against Existing Models / Methods using Superfamily Dataset

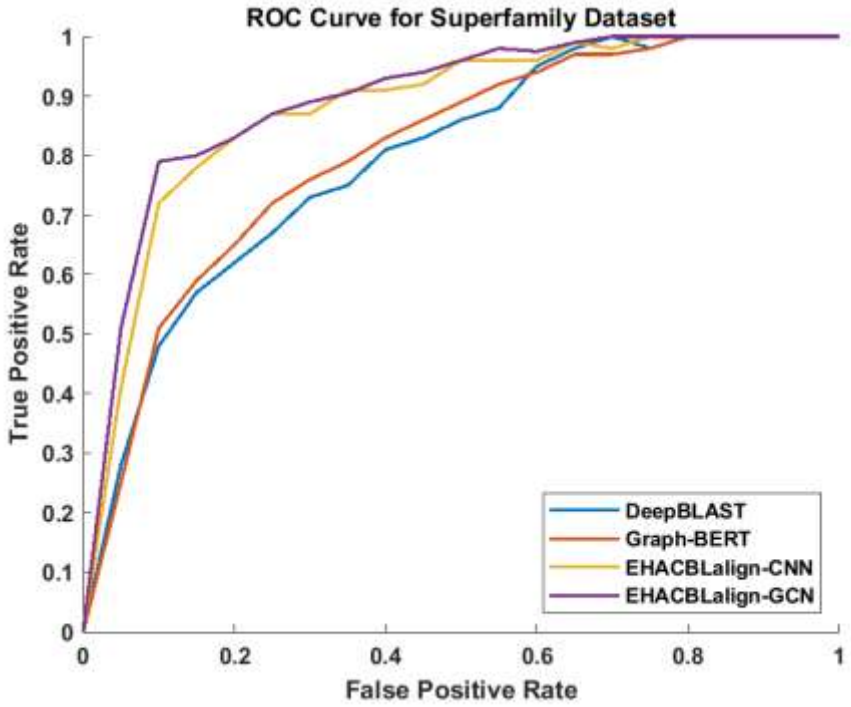
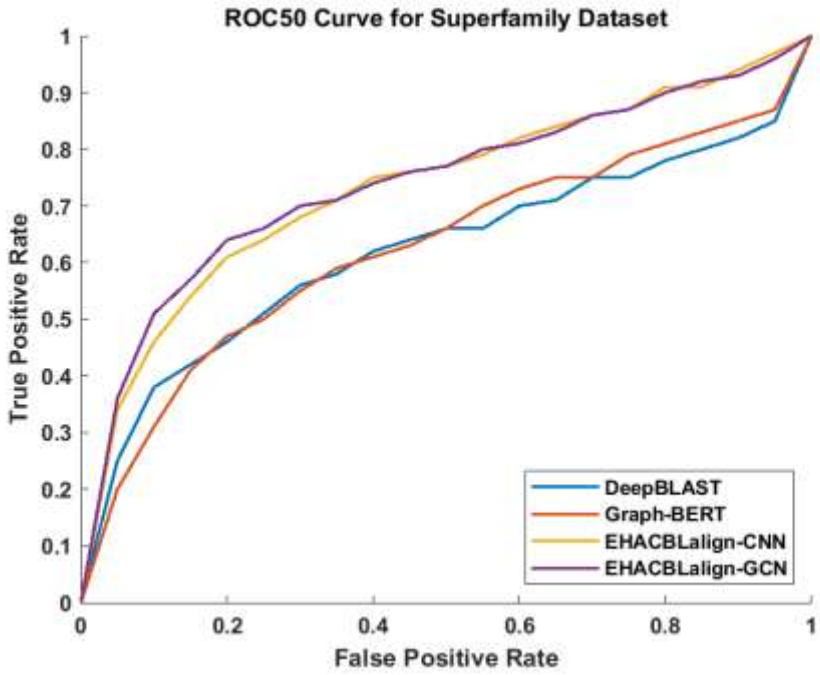


Figure 8(b). Comparison of ROC50 for EHACBLalign-GCN Method against Existing Models / Methods using Superfamily Dataset



Conclusion

In this work, the EHACBLalign-GCN method was developed to improve the evaluation metrics of detecting protein remote homologies. In this method, the EHACBLalign-GCN prunes uninformative edges for better prediction of accuracy, precision, recall and F-measure. EHACBLalign-GCN method is trained using the ground truth labels for predicting families. The experimental results proved that the EHACBLalign-GCN method achieved better values for the evaluation metrics viz., accuracy, precision, recall and F-measure, with regard to Remote Homology Detection and Fold Recognition.

References

1. Altschul S. F., Gish W., Miller W., Myers E. W., Lipman, D. J. Basic local alignment search tool. Journal of Molecular Biology, 215(3), 1990, 403-410
2. Gopinath K., Rajendran G. Exploring the Potential of Deep Learning in Protein Remote Homology Detection and Folds Identification using Transfer Learning and Attention Mechanism. IRE Journals, 7(3), 2023, 16-23
3. Eddy S. R. Profile hidden Markov models. Bioinformatics, 14(9), 1998, 755-763
4. Chothia C., Lesk A. M. The relation between the divergence of sequence and structure in proteins. EMBO Journal, 5(4), 1986, 823-826
5. Liao L., Noble W. S. Combining pairwise sequence similarity and support vector machines for remote protein homology detection. Bioinformatics, 19(13), 2003, 1523-1530
6. Hochreiter S., Schmidhuber, J. Long short-term memory. Neural Computation, 9(8), 1997, 1735-1780
7. Wang Z., Xu J., Yu Y. Protein remote homology detection using a convolutional neural network. Bioinformatics, 27(12), 2011, 1696-1702
8. Weston J., Leslie C., Ie E., Zhou D., Elisseeff A., Noble W. S. Semi-supervised protein classification using cluster kernels. Bioinformatics, 21(15), 2004, 3241-3247
9. Zhang S., Zhou, S. Finding related genes based on protein-protein interaction networks. BMC Bioinformatics, 14(1), 2013, 271

10. Paccanaro A., Casbon J. A., Saqi, M. A. Spectral clustering of protein sequences. Nucleic Acids Research, 34(5), 2006, 1571-1580
11. Jiang Y., Xu, D. Protein structure comparison: Algorithms and applications. International Journal of Bioinformatics Research and Applications, 3(4), 2007, 402-425
12. Gopinath K., Rajendran G. HACBLALIGN: A Hierarchal Attention Based Deep Learning Framework for Protein Remote Homology Detection and Fold Identification. Journal of Theoretical and Applied Information Technology, 104 (14) 2023, 5578 – 5588
13. Kingma D. P., Ba J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR). 2014
14. Snel B., Lehmann G., Bork P., Huynen M. A. STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Research, 28(18), 2000, 3442-3444
15. Hamilton W. L., Ying R., Leskovec, J. Inductive representation learning on large graphs. In Proceedings of the 31st International Conference on Neural Information Processing Systems, 2000, 1024-1034
16. Grover A., Leskovec J. Node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, 855-864
17. Mikolov T., Chen K., Corrado G., Dean, J. Efficient estimation of word representations in vector space. In Proceedings of the International Conference on Learning Representations (ICLR). 2013.
18. Rao R. M., Liu J., Verkuil R., Meier J., Canny J., Abbeel P., Sercu, T. MSA transformer. In Proceedings of the 38th International Conference on Machine Learning, 2021, 8844-8856
19. Velickovic P., Cucurull G., Casanova A., Romero A., Lio P., Bengio Y. Graph attention networks. In International Conference on Learning Representations (ICLR), 2018.
20. Chen J., Yuan L., Zhang Y., Lu H. Graph convolutional networks for molecular property prediction with graph attention mechanism. Journal of Chemical Information and Modeling, 60(4), 2020, 1919-1929
21. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 770-778